Chromosome Contact Matrices: What, how and why?

Bartek Wilczyński
bartek@mimuw.edu.pl
http://regulomics.mimuw.edu.pl/

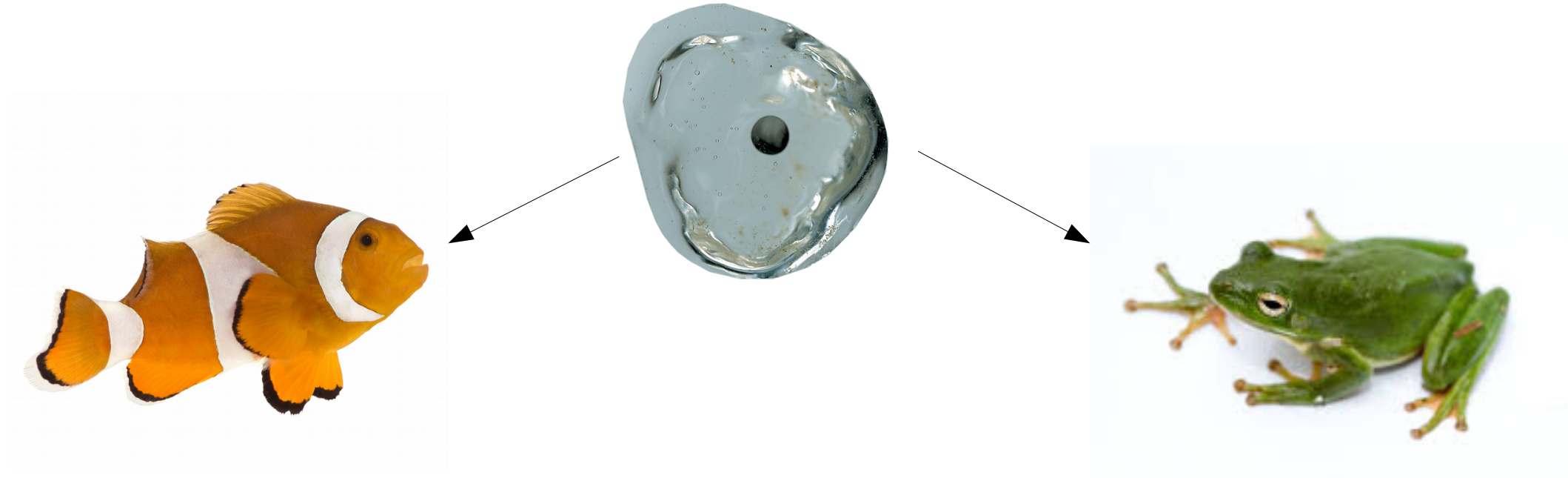
University of Warsaw
Faculty of Mathematics,
Informatics and Mechanics

Faculty colloquium

Warsaw June 8th 2017

XIX century question: where is the information?

- How does a cell know what to become?
- How does it know how big the organism should be?
- How does it know that the child should be similar to its parents?



THE CHEMICAL BASIS OF MORPHOGENESIS

By A. M. TURING, F.R.S. *University of Manchester*

(*Received* 9 *November* 1951—*Revised* 15 *March* 1952)

It is suggested that a system of chemical substances, called morphogens, reacting together and diffusing through a tissue, is adequate to account for the main phenomena of morphogenesis. Such a system, although it may originally be quite homogeneous, may later develop a pattern or structure due to an instability of the homogeneous equilibrium, which is triggered off by random disturbances. Such reaction-diffusion systems are considered in some detail in the case of an isolated ring of cells, a mathematically convenient, though biologically unusual system. The investigation is chiefly concerned with the onset of instability. It is found that there are six essentially different forms which this may take. In the most interesting form stationary waves appear on the ring. It is suggested that this might account, for instance, for the tentacle patterns on *Hydra* and for whorled leaves. A system of reactions and diffusion on a sphere is also considered. Such a system appears to account for gastrulation. Another reaction system in two dimensions gives rise to patterns reminiscent of dappling. It is also suggested that stationary waves in two dimensions could account for the phenomena of phyllotaxis.

The purpose of this paper is to discuss a possible mechanism by which the genes of a zygote may determine the anatomical structure of the resulting organism. The theory does not make any new hypotheses; it merely suggests that certain well-known physical laws are sufficient to account for many of the facts. The full understanding of the paper requires a good knowledge of mathematics, some biology, and some elementary chemistry. Since readers cannot be expected to be experts in all of these subjects, a number of elementary facts are explained, which can be found in text-books, but whose omission would make the paper difficult reading.

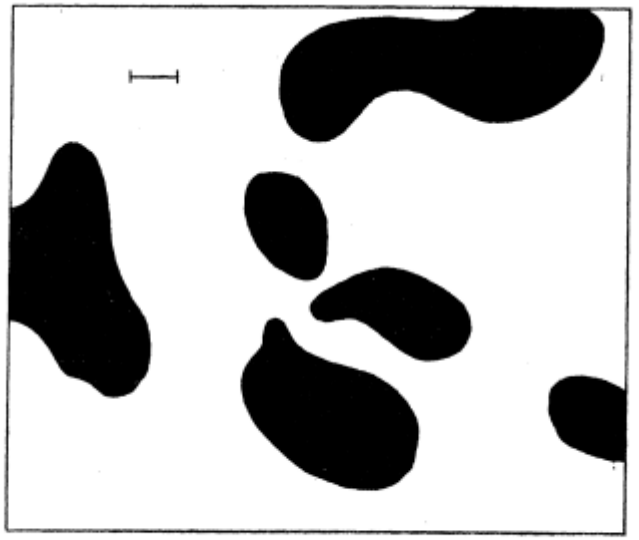
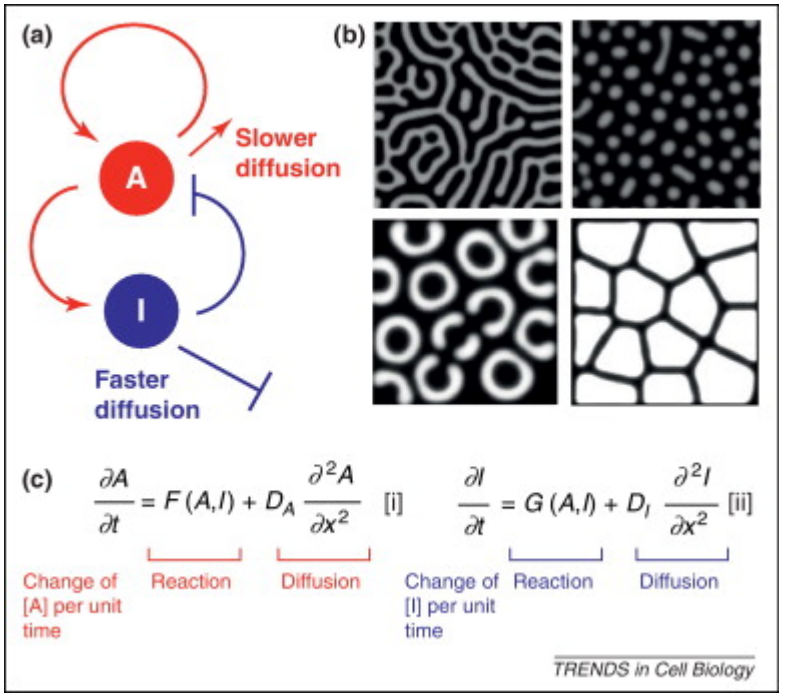
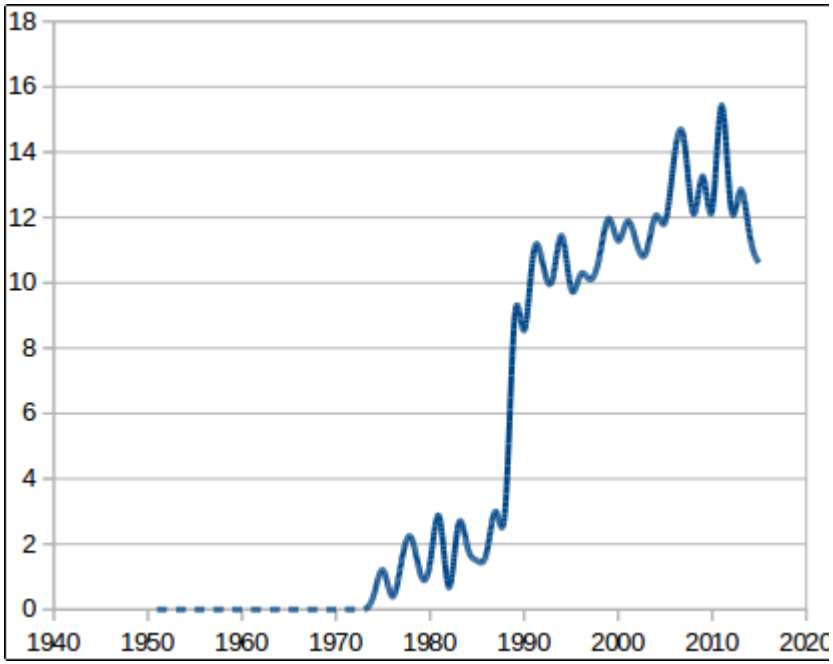


FIGURE 2. An example of a 'dappled' pattern as resulting from a type (*a*) morphogen system. A marker of unit length is shown. See text, §9, 11.

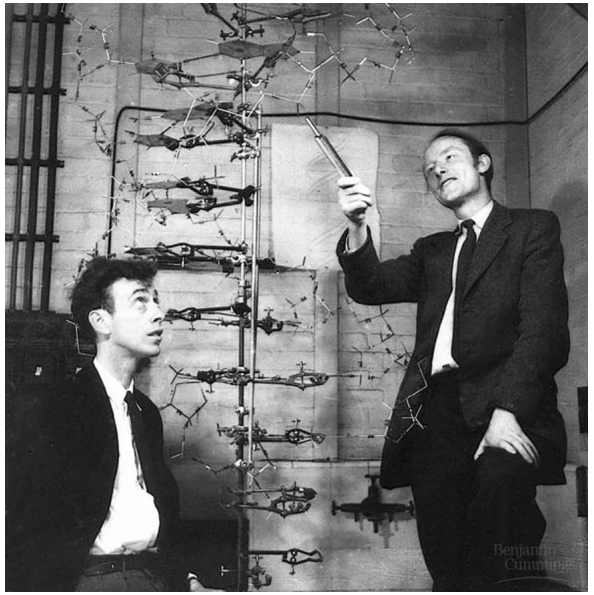




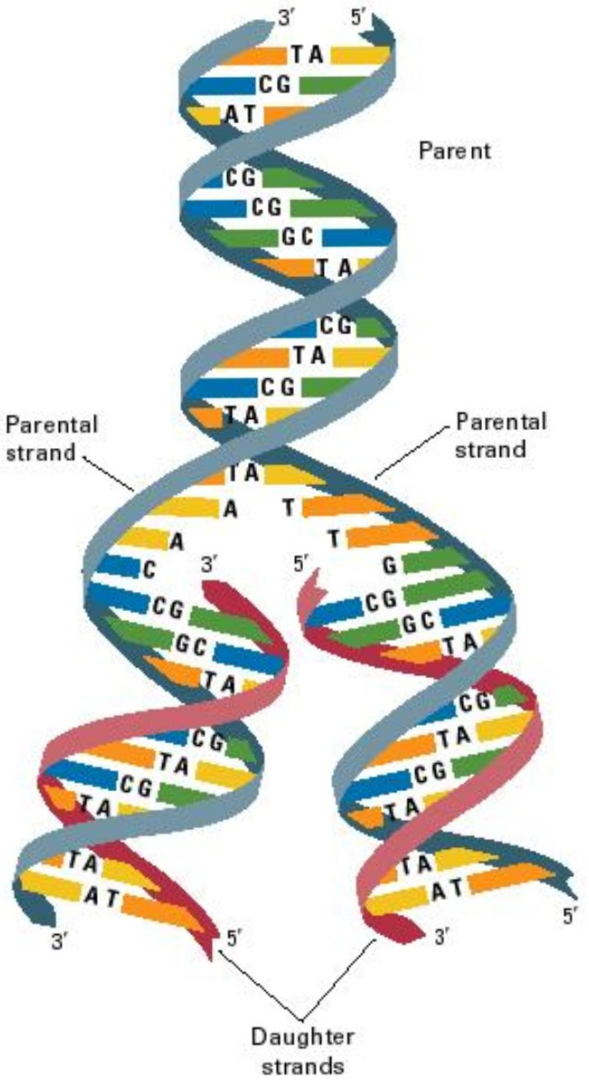
"morphogen" in publications

Understanding the information storage

- The hereditary information is kept in DNA
- It is a discrete code, that is copied exactly

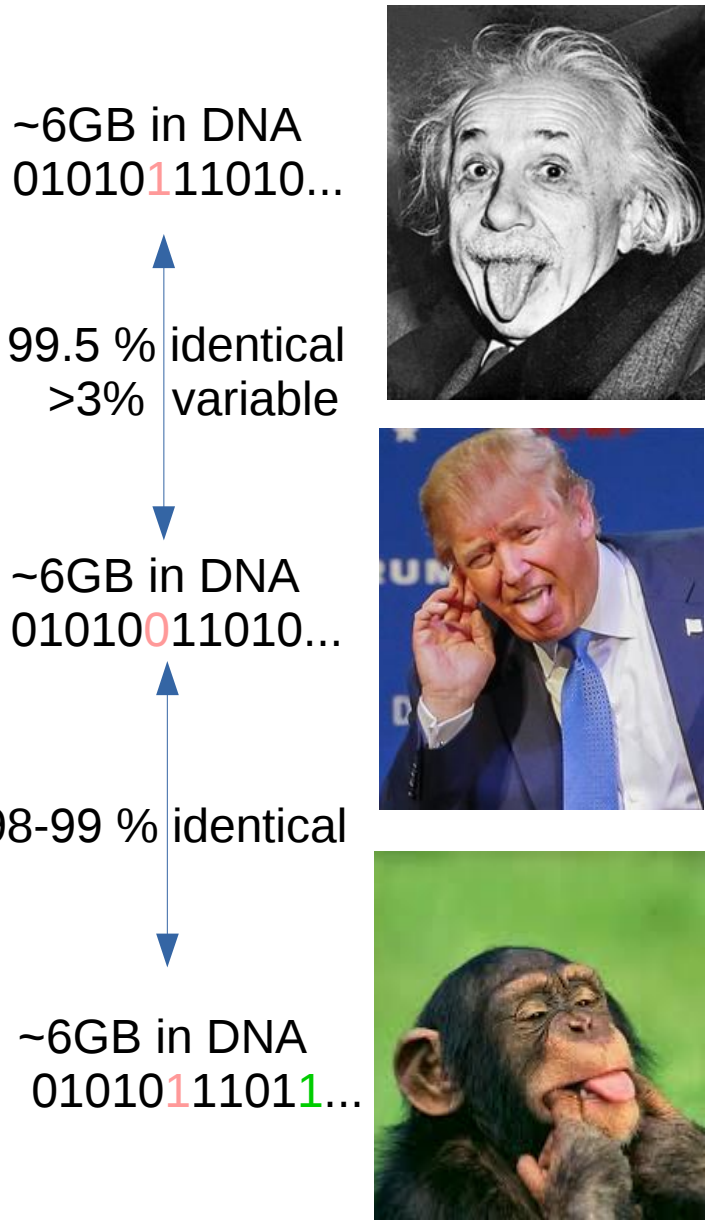


Watson&Crick 1953

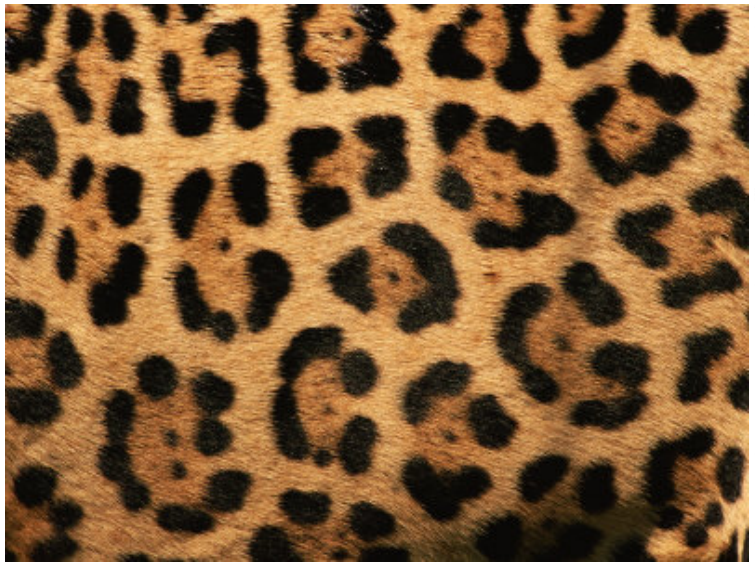


XX century question: how is it encoded in DNA?

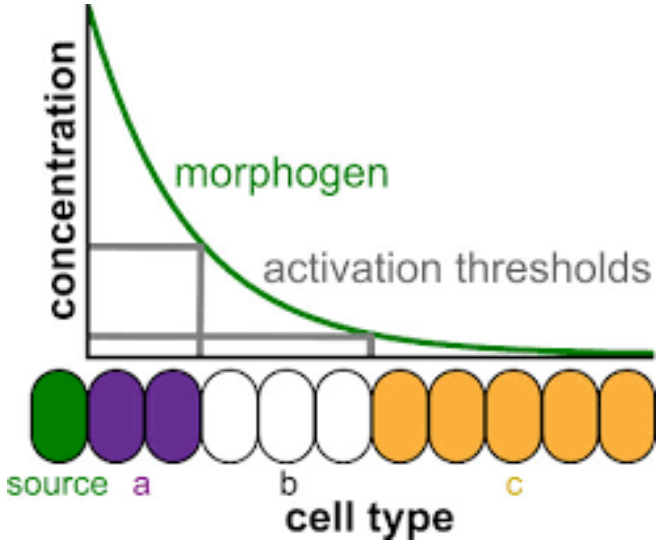
- We now know thousands of genome DNA sequences
- We can do statistics on variants and use it in medicine, forensics, paternity testing, etc.
- We know very little about mechanisms of translating DNA variants into phenotype



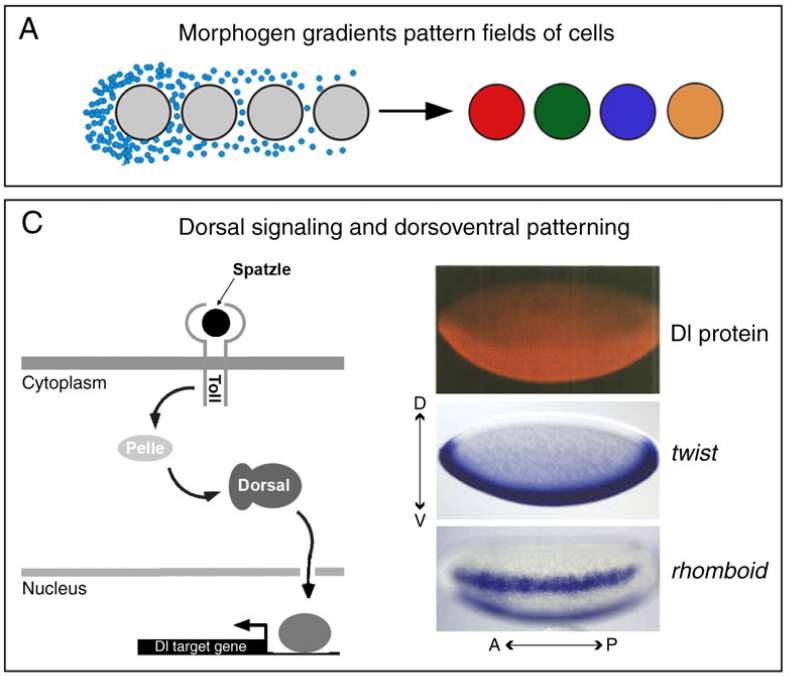
Back to Alan Turing's morphogens

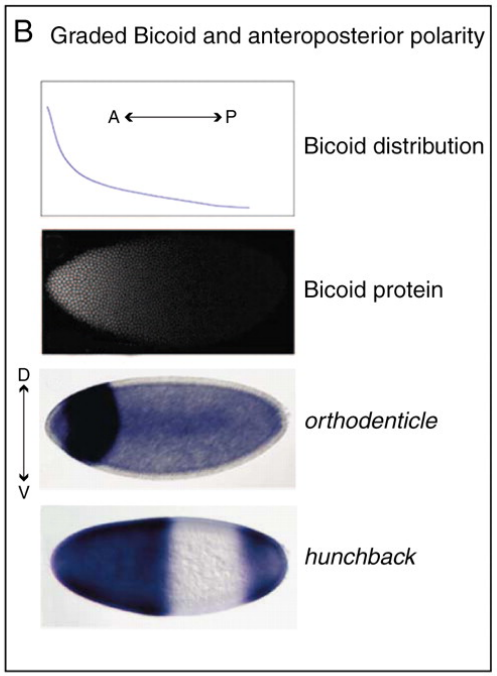


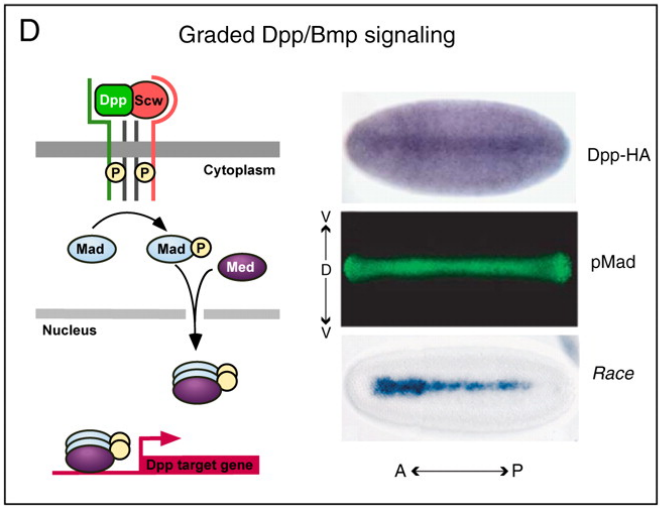
Skin pattern

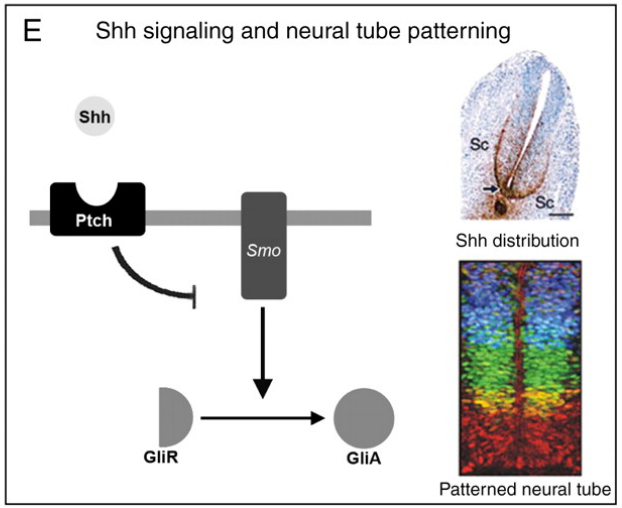


Molecular level

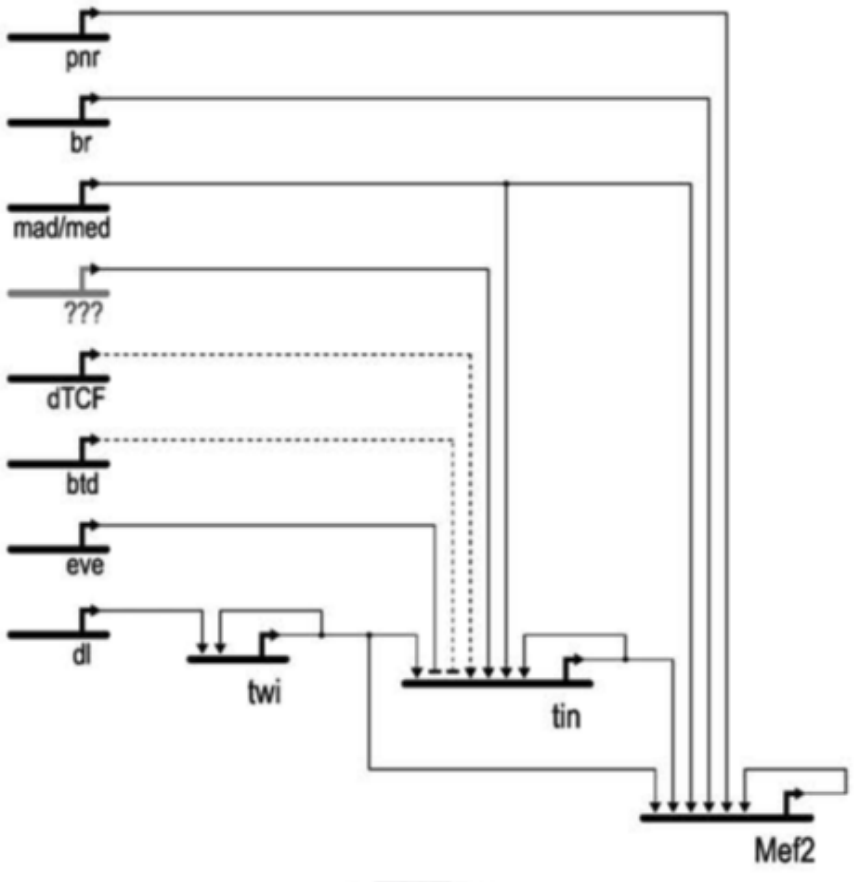


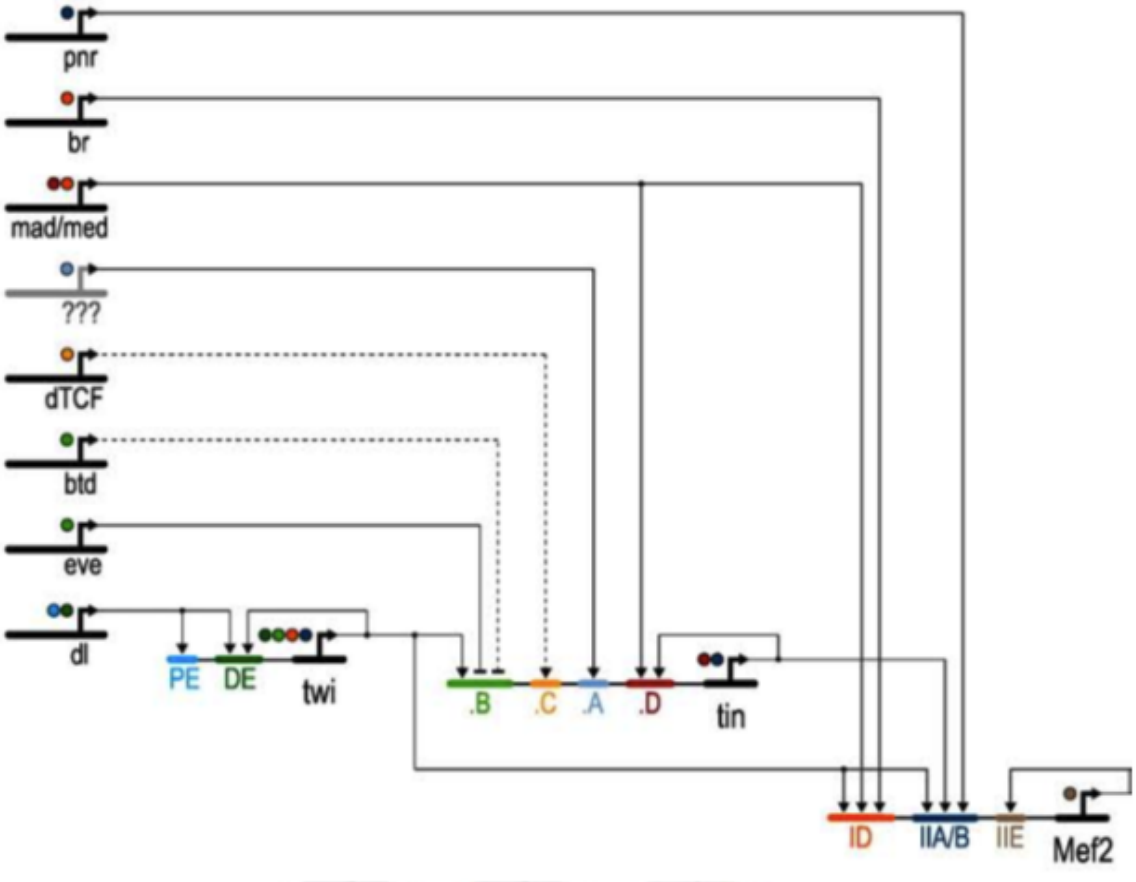


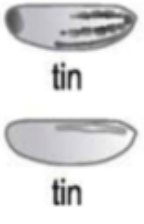




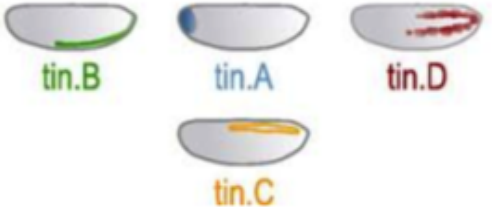
Gene regulatory networks



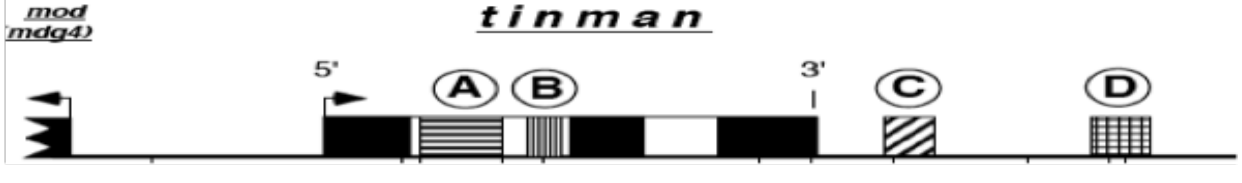




A

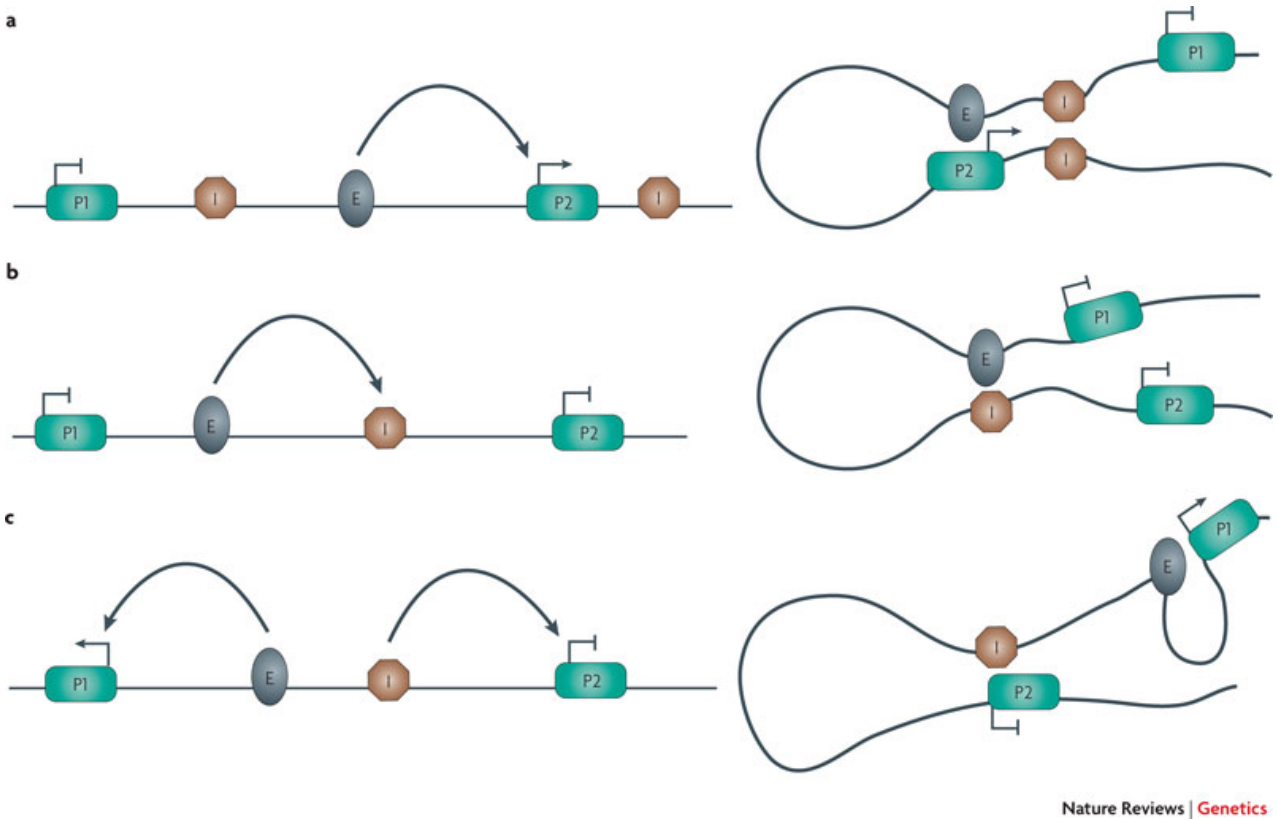


B

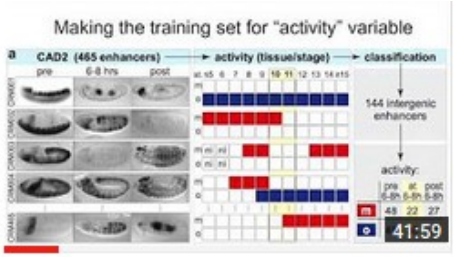


Wilczynski & Furlong, Dev. Biol, 2010

Regulatory genome – parts list



- ~5-10k insulators
- ~ 20k genes
- ~100k start sites
- ~ 1 milion regulatory elements



Bartek Wilczynski - Using Python to Find a Bayesian Network Describing Your Data

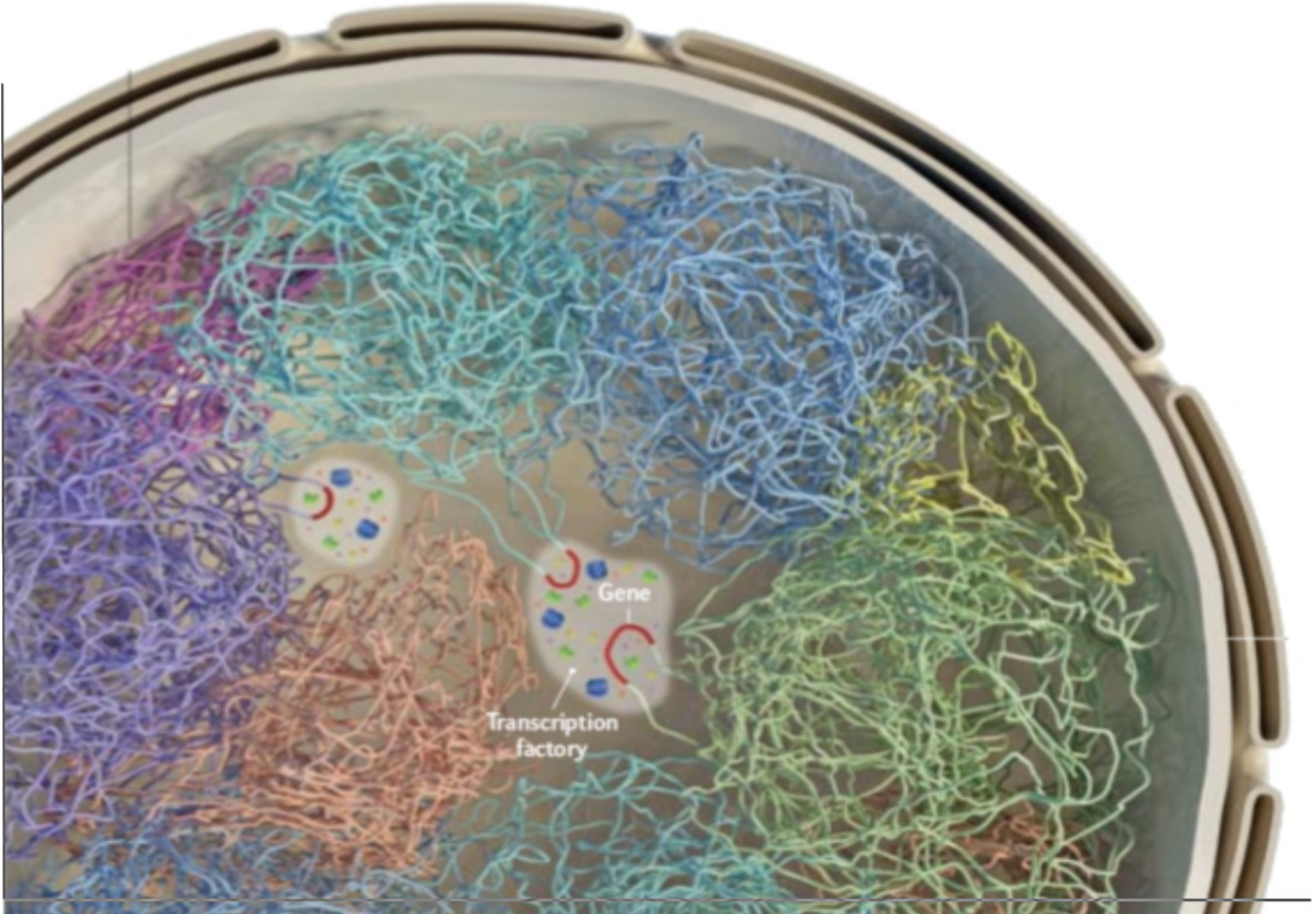
PyData
2 years ago • 5,670 views

PyData SV 2014 Today's world is full of data that is easily accessible for anyone. The problem now is how to make sense of this ...

http://youtu.be/WD7b1jLeZZ4

Patelak et al BMC Bioinf. 2009
BW&Furlong MSB 2010
Bonn et al. Nature Gen. 2012
Podsiadlo et al BMC SysBio 2013
Bednarz&BW JBCB 2014
BW&Tiuryn JCB 2017
Herman-Izycka et al BMC Med. Gen. 2017

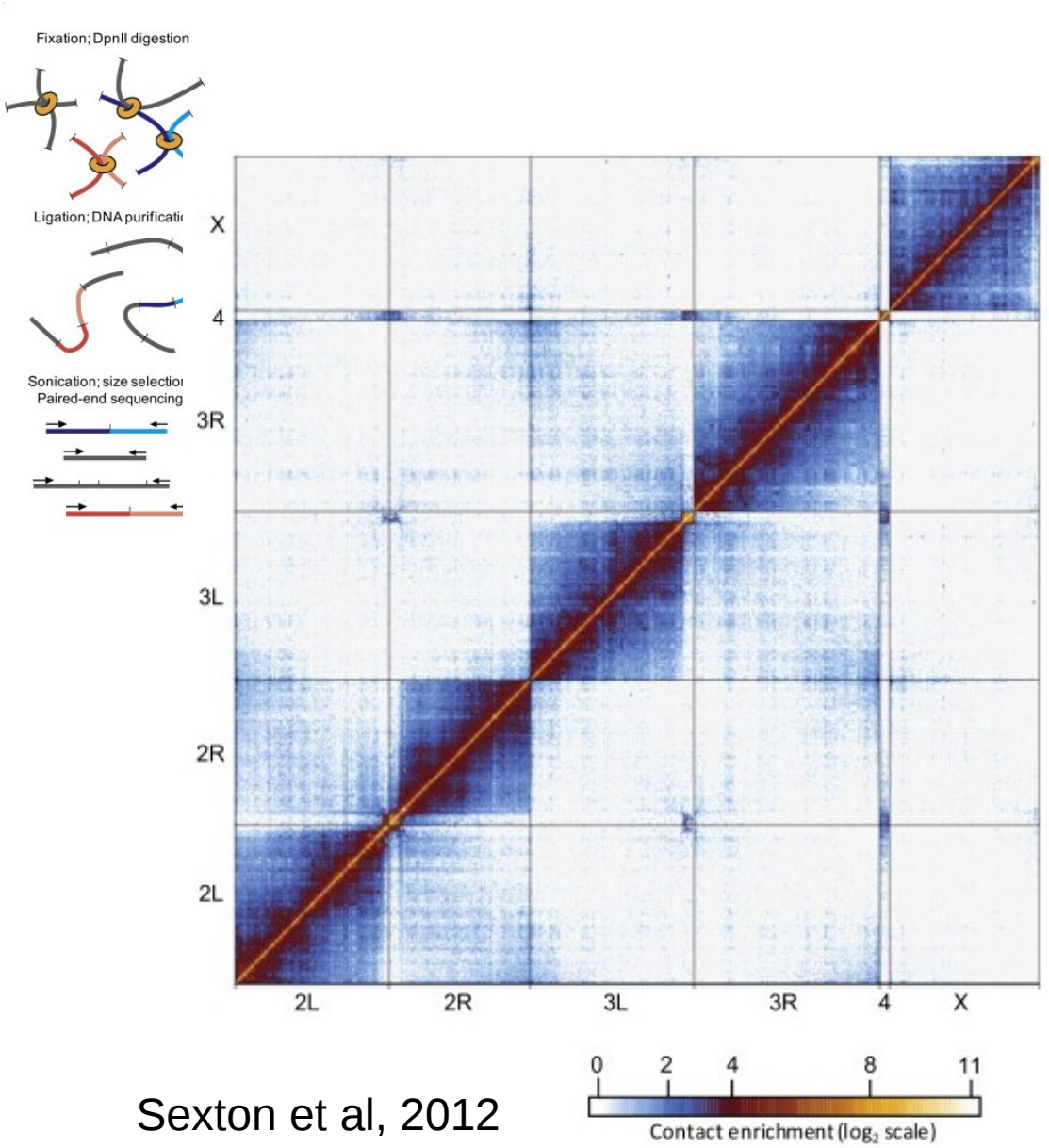
All of this happens in 3D in a dense environment



Too small to see the actual contacts, very dynamic structures

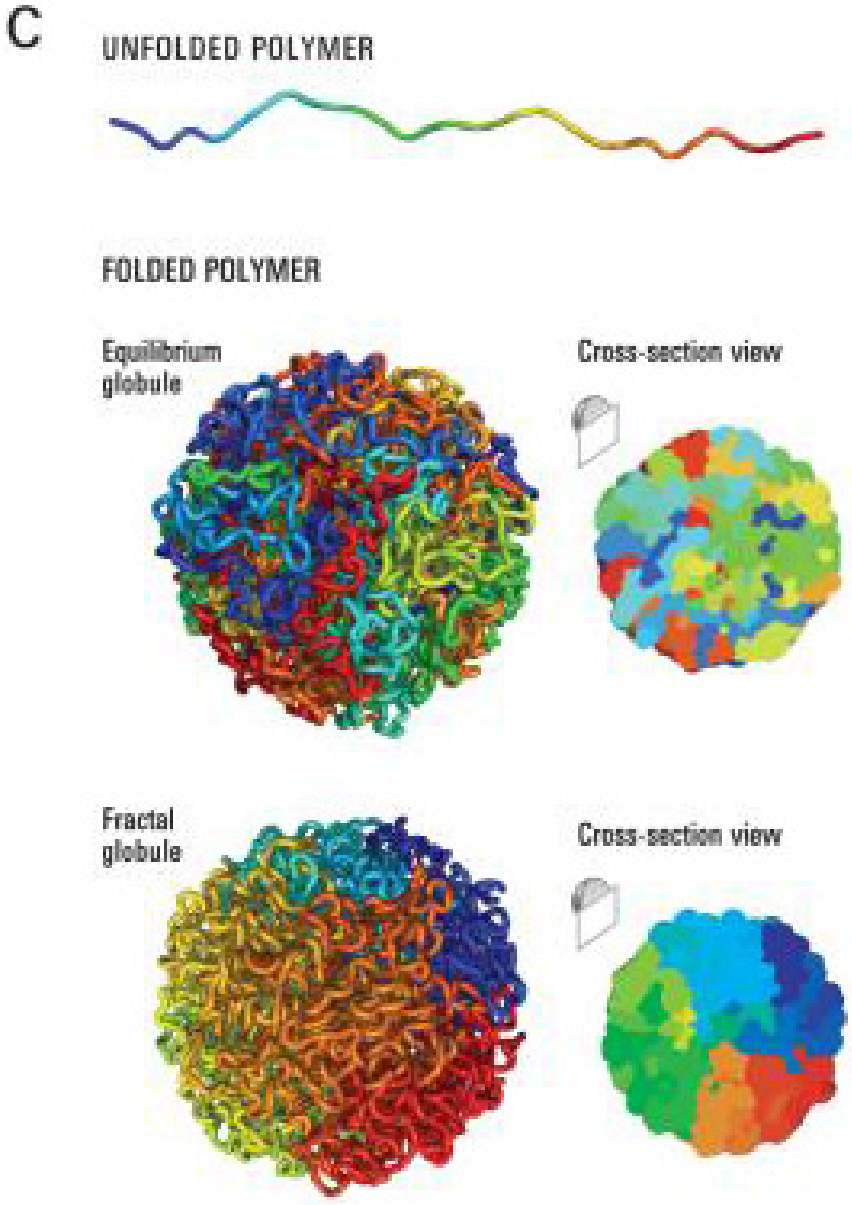
High-throughput Chromosome Conformation capture (Hi-C)

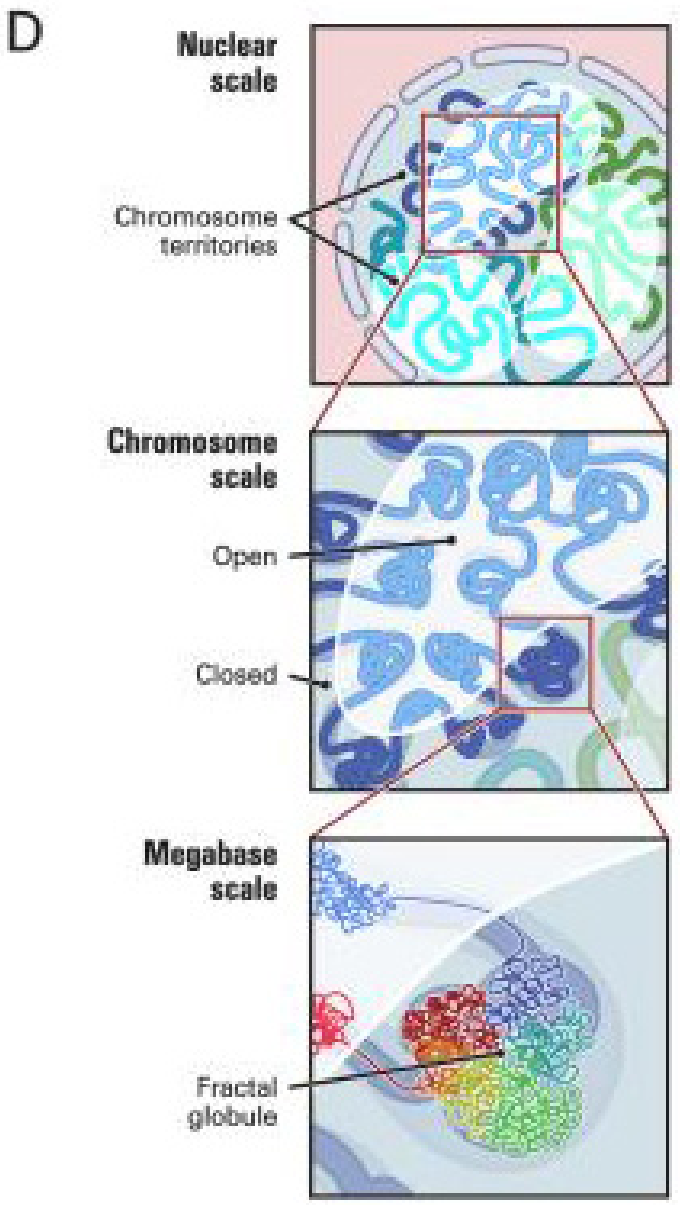
- We can use biochemical protocol to identify contacts between different DNA fragments
- This gives us a large square, symmetric, positive matrix, where each entry describes the number of observed contacts between fragments *i* and *j*



Sexton et al, 2012

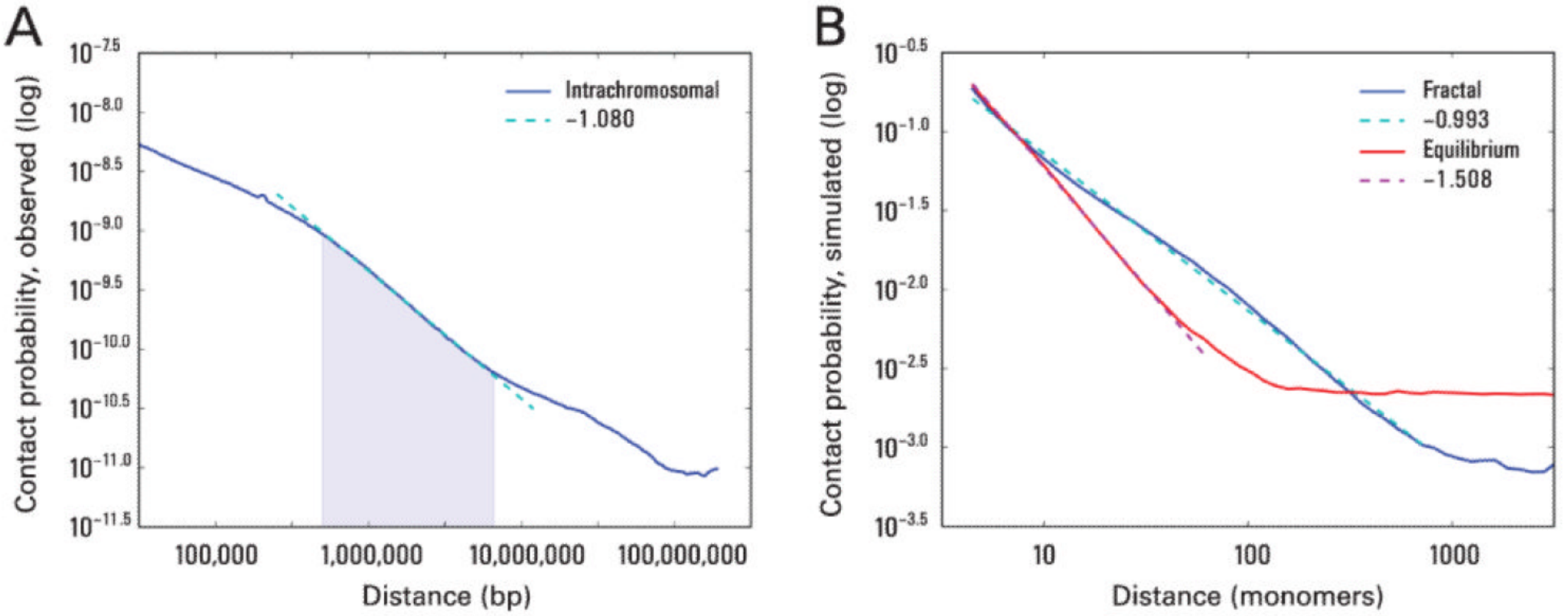
Different kinds of structures – different distributions





Lieberman Aiden et al, Science 2009

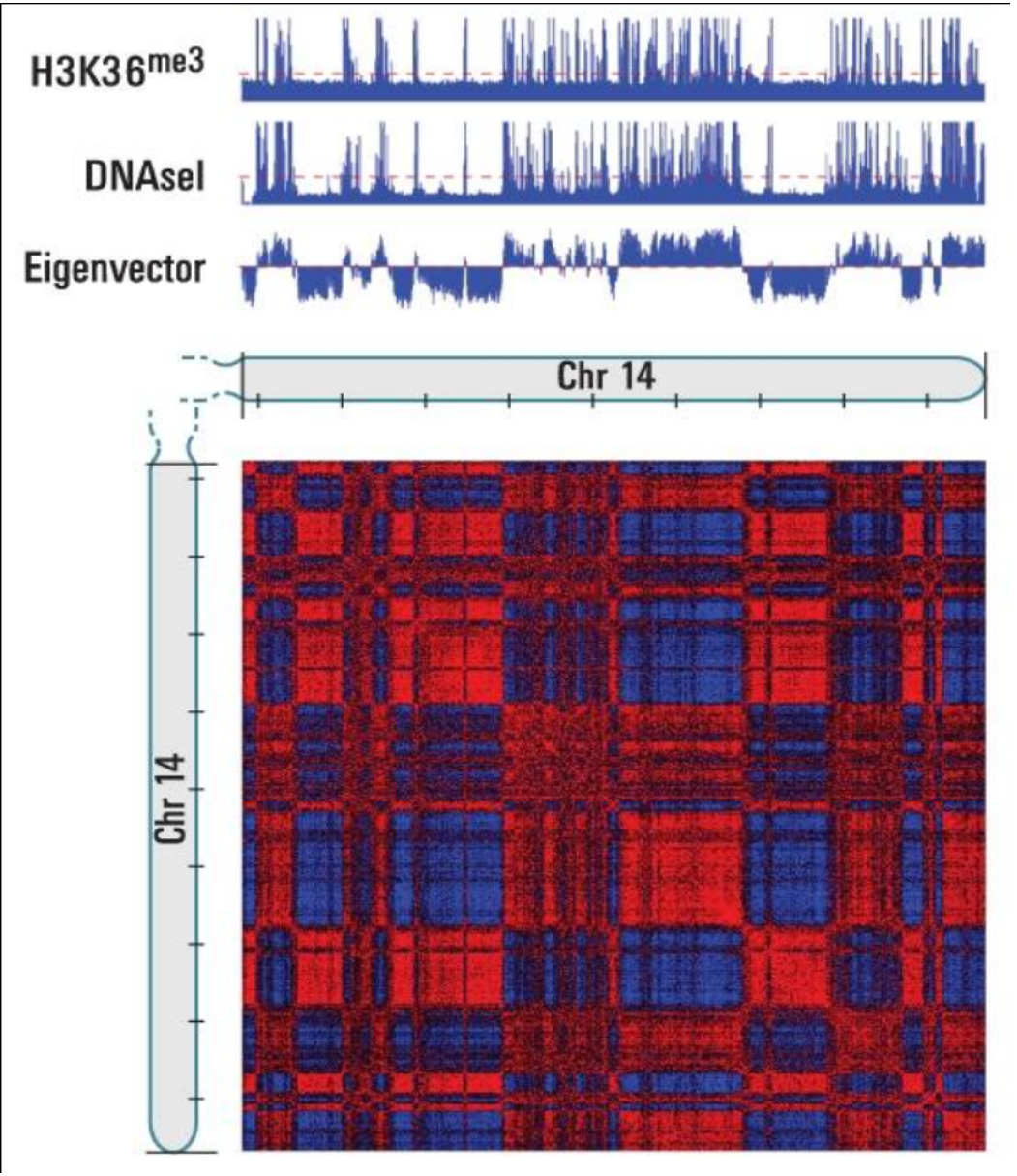
Approximating real data by Power-law distributions



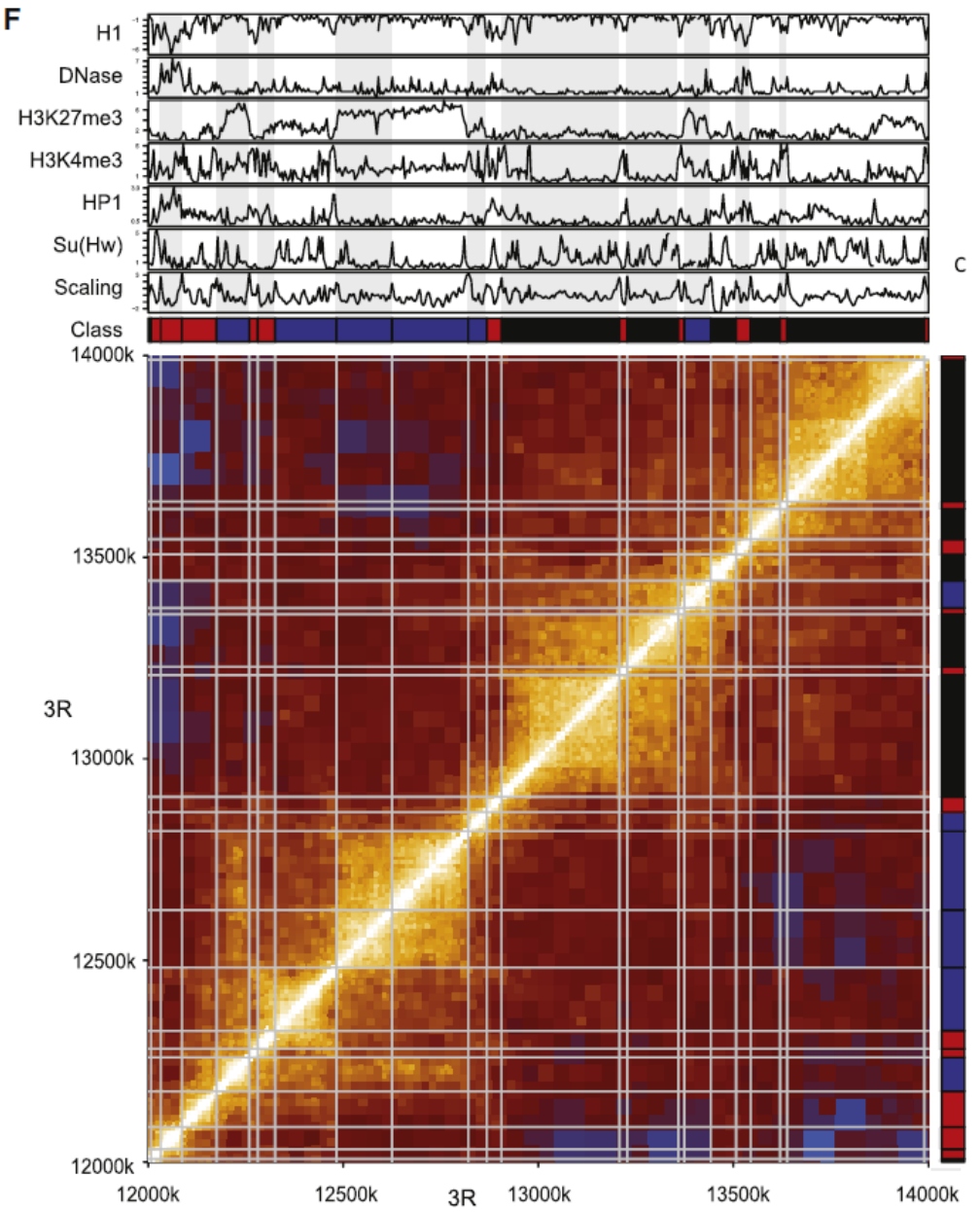
It's a fractal globule !

Lieberman Aiden et al, Science 2009

A/B compartments vs Topologically Associating Domains



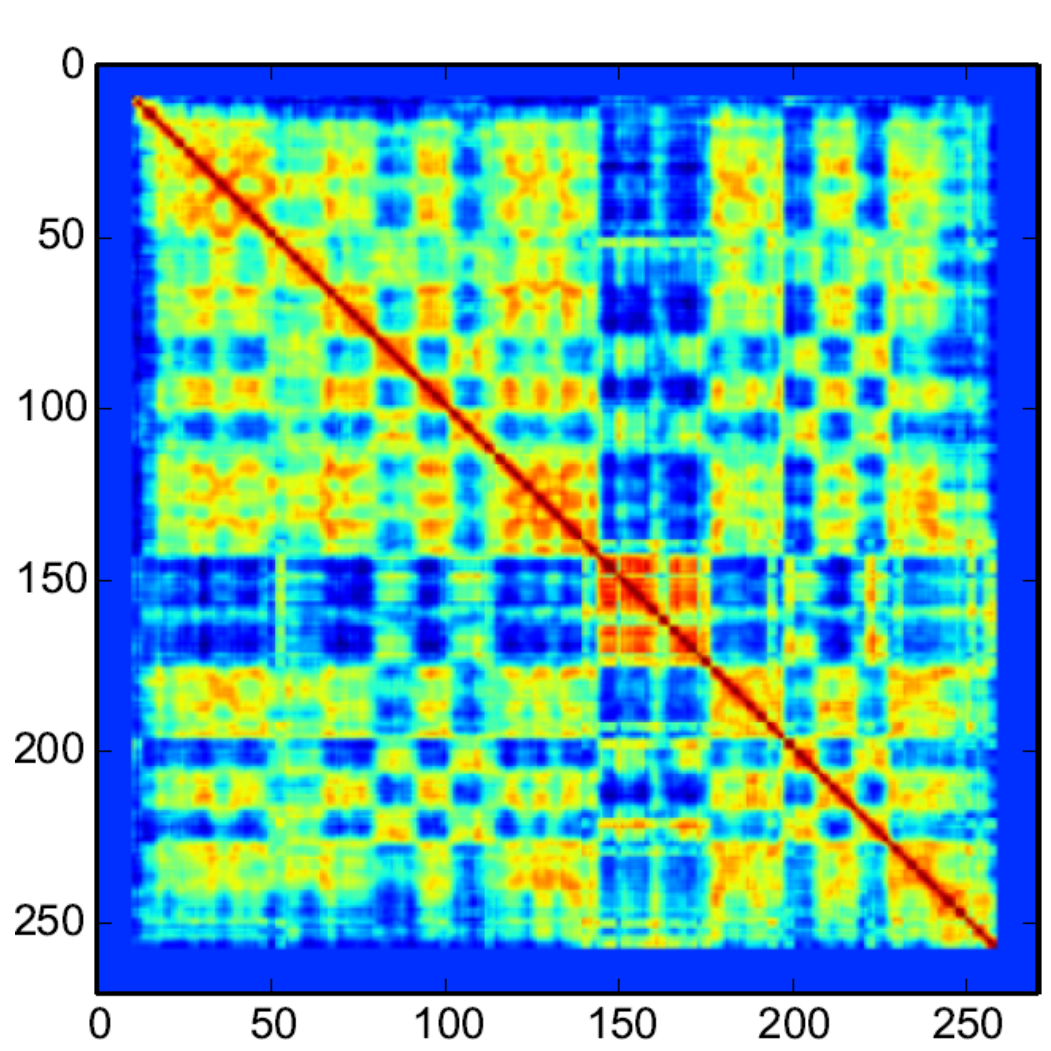
Lieberman-Aiden et al, Nature, 2009

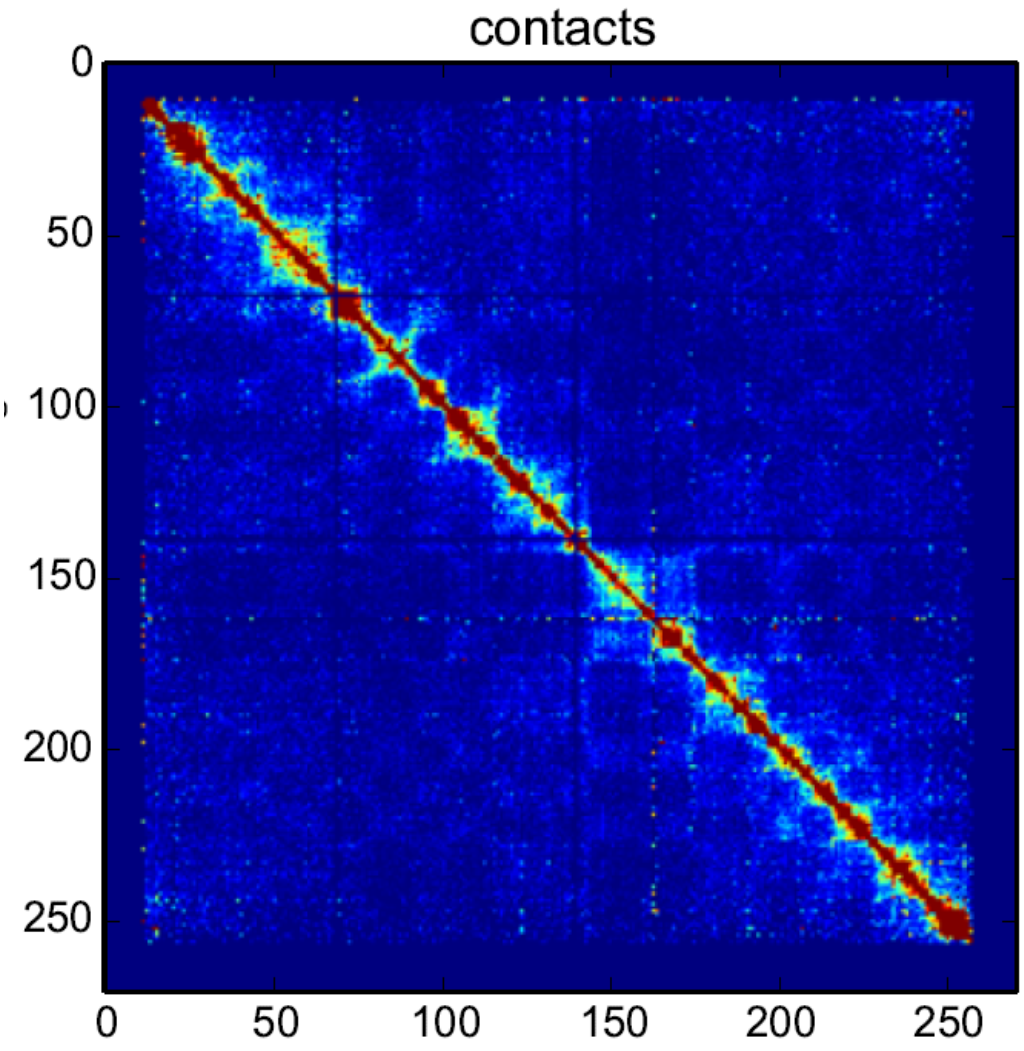


Sexton et al, Cell, 2012

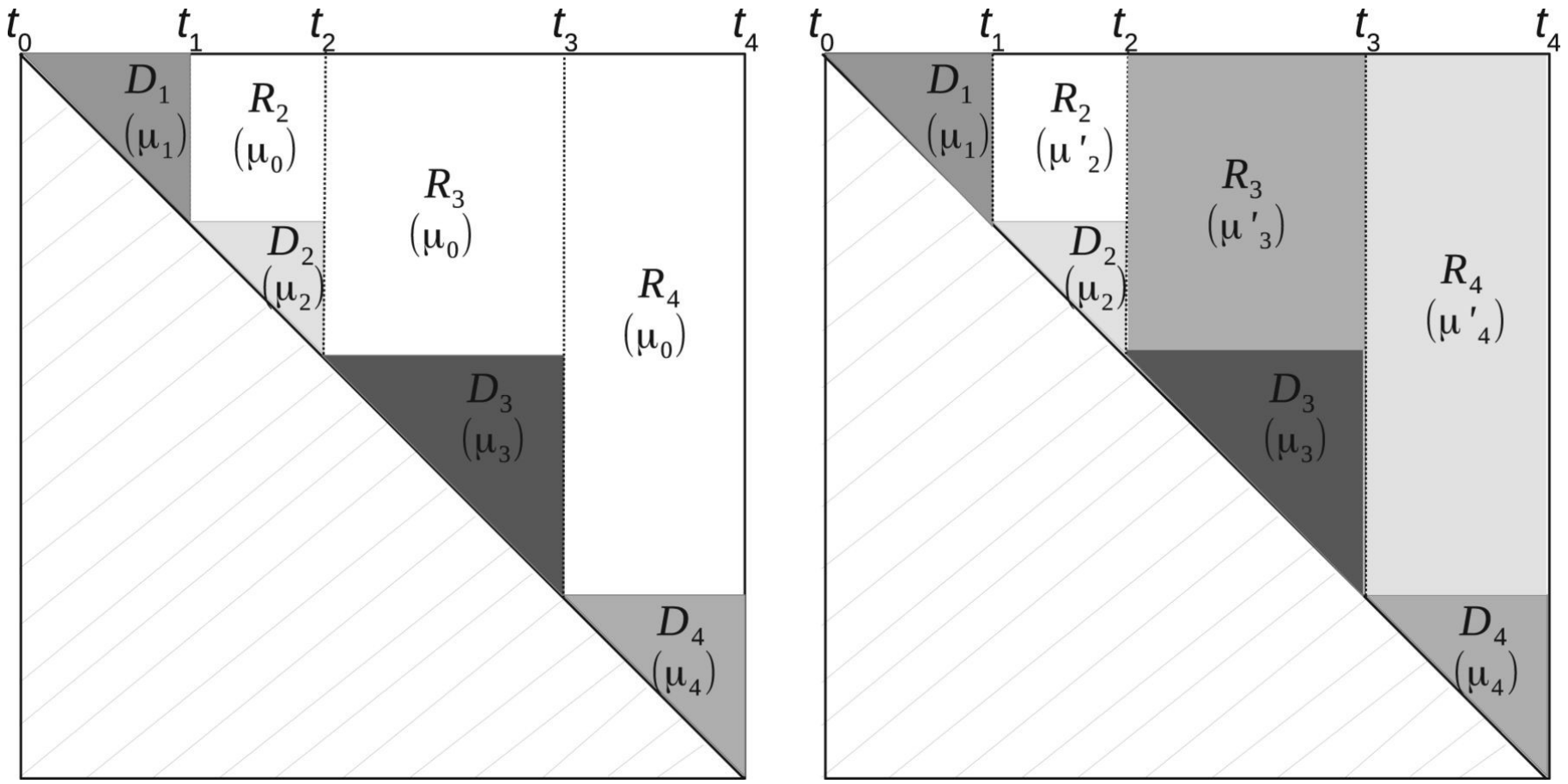
Depending on processing of the data, TADs or compartments are more visible.

While compartments tend to be larger than TADs, they do overlap in terms of scale and some TADs span more than one compartment.





Examples of block diagonal and extended block diagonal matrices (μi,j)1≤i≤j≤n.

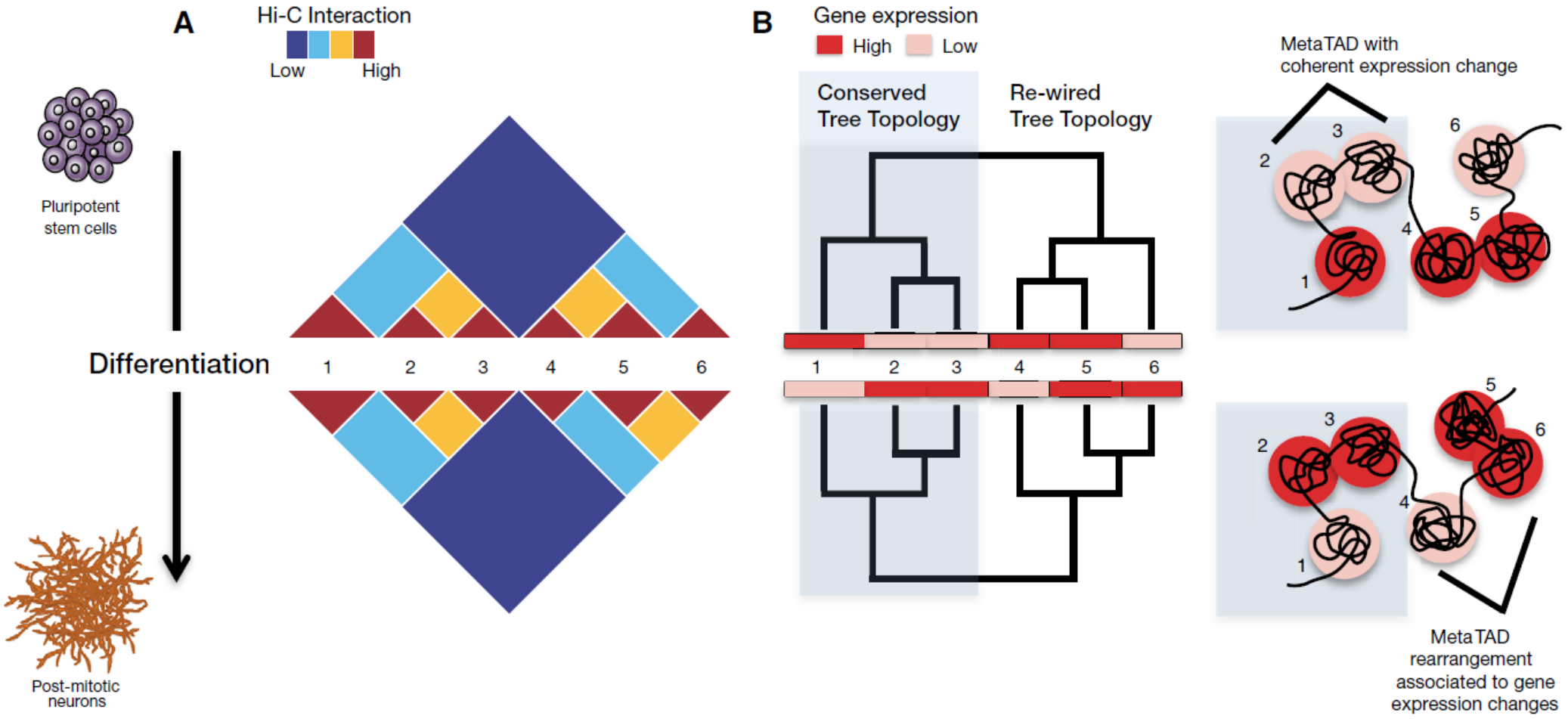


Celine Lévy-Leduc et al. Bioinformatics 2014;30:i386-i392

© The Author 2014. Published by Oxford University Press.

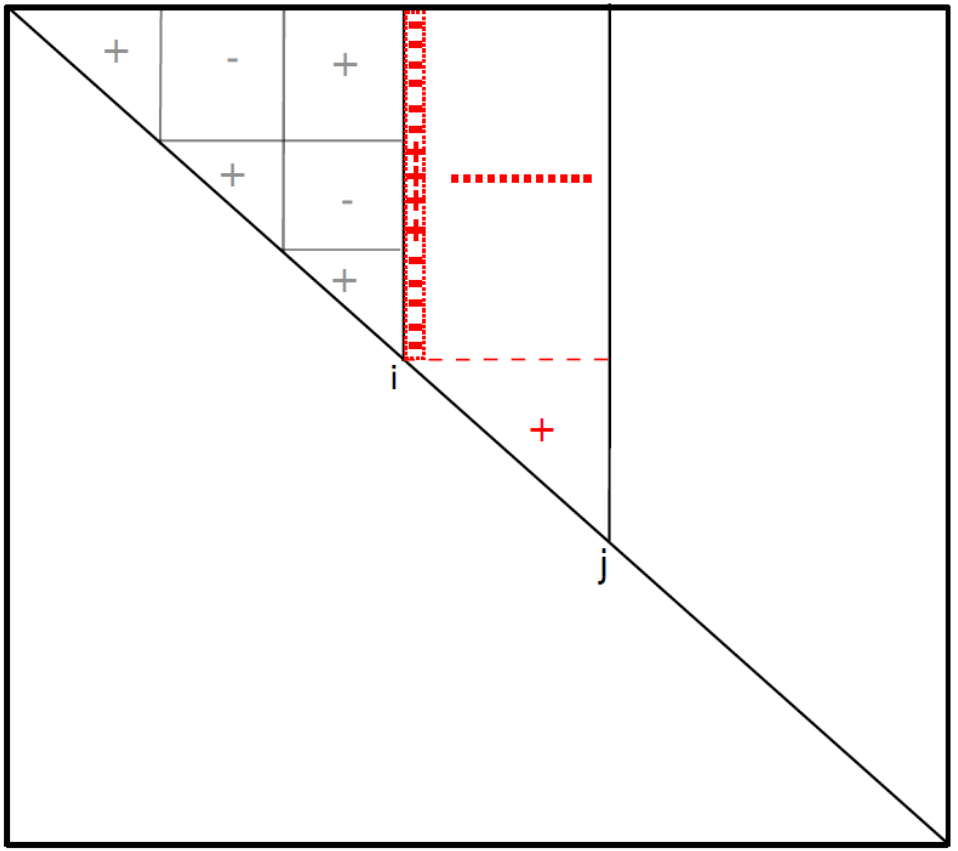
Bioinformatics

Aggregation of TADs by their contact frequency



Fraser et al, Mol. Sys. Bio. 2015

Top-down approach for hierarchical compartment delineation - OPPA



- Start with the correlation matrix for a given chromosome
- Using dynamic programming find the optimal segmentation into „opposite” compartments
- The dynamic algorithm will find the global optimum, provided that the data is consistent with the compartment model
- For every compartment run the method recursively, on re-normalized sub-matrix
- The process naturally stops when all vectors in the compartment have positive correlation

Works in polynomial time, while other approaches to optimal solutions require exponential time

Wilczynski, unpublished

Hierarchical Regulatory Domain Inference from Hi-C Data

Simons Institute

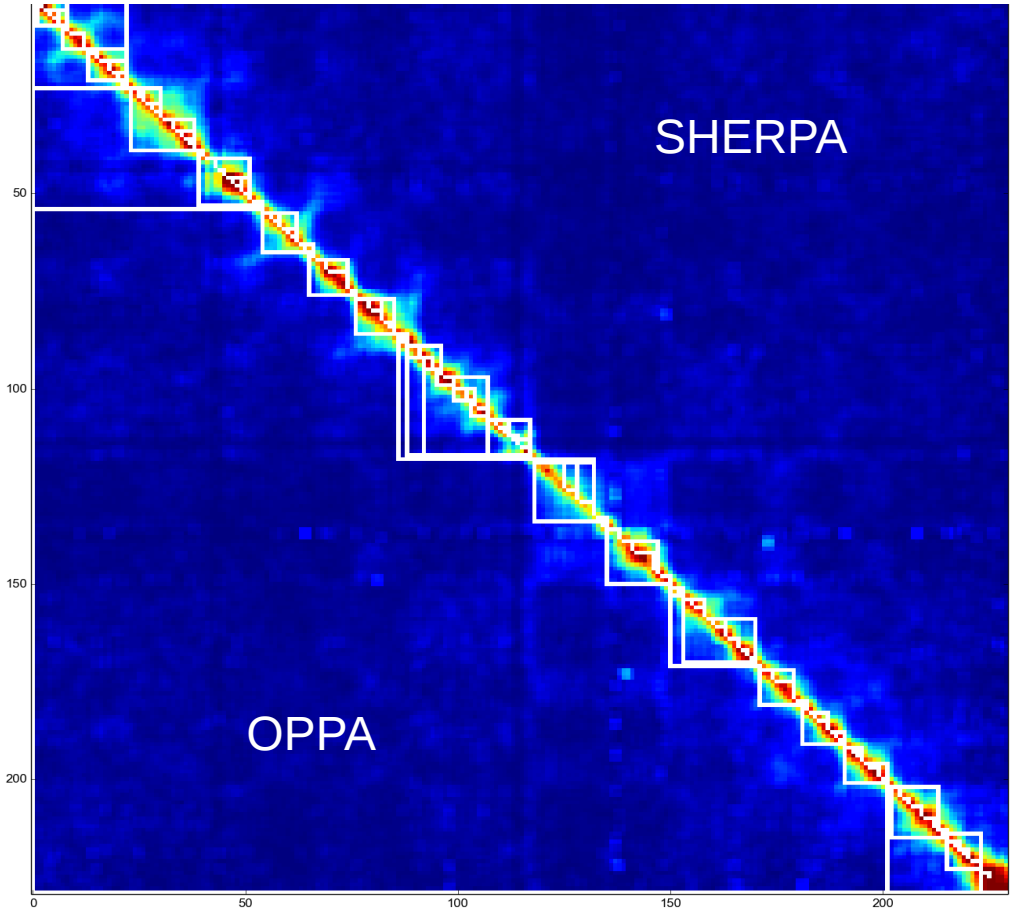
1 year ago • 149 views

Bartek **Wilczynski**, University of Warsaw Regulatory Genomics and Epigenomics ...

http://youtu.be/pxsz6QOtzHA

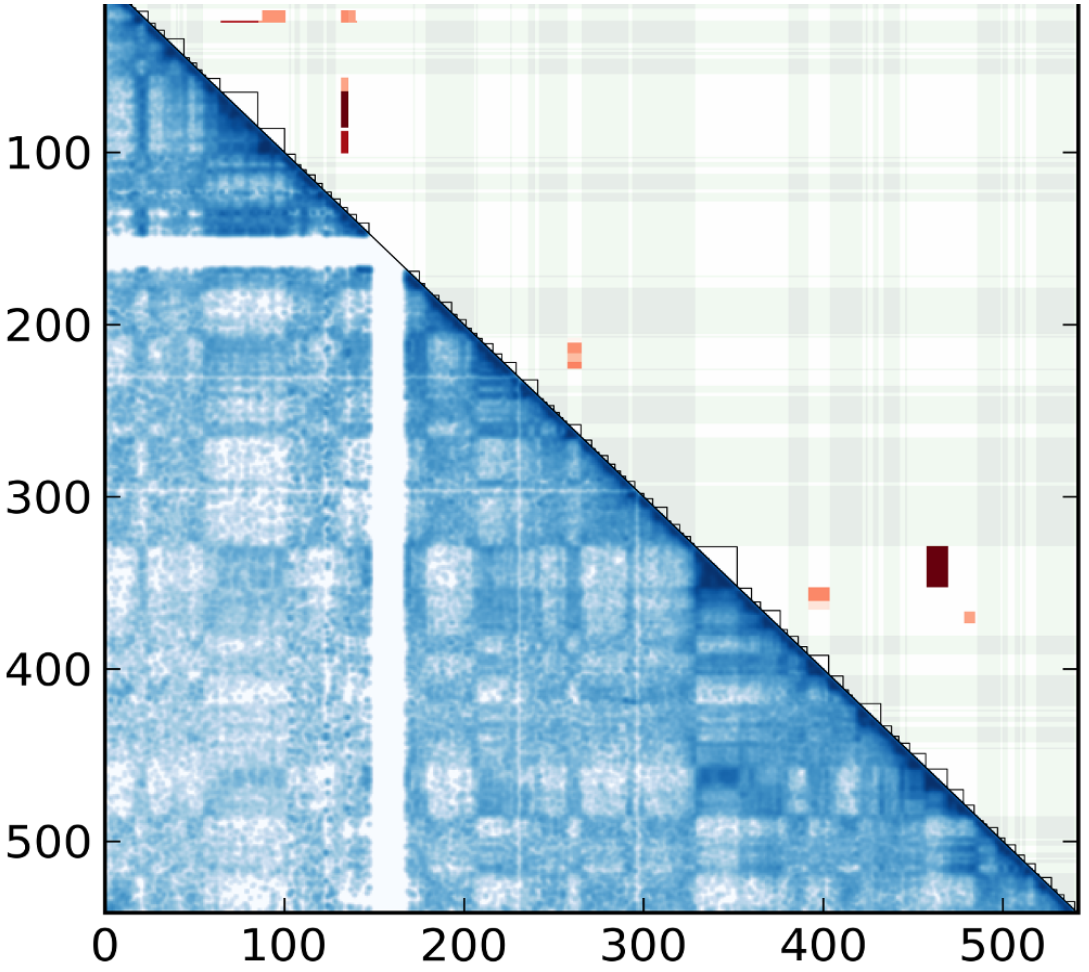
Comparing SHERPA to OPPA

- Both methods have natural means of stopping the segmentation
  - SHERPA stops aggregation when there is no gain over simpler model
  - OPPA stops recursion, when all profiles are positively correlated
- They tend to stop at the same level, however sometimes their scales do overlap



Long range-contacts identification

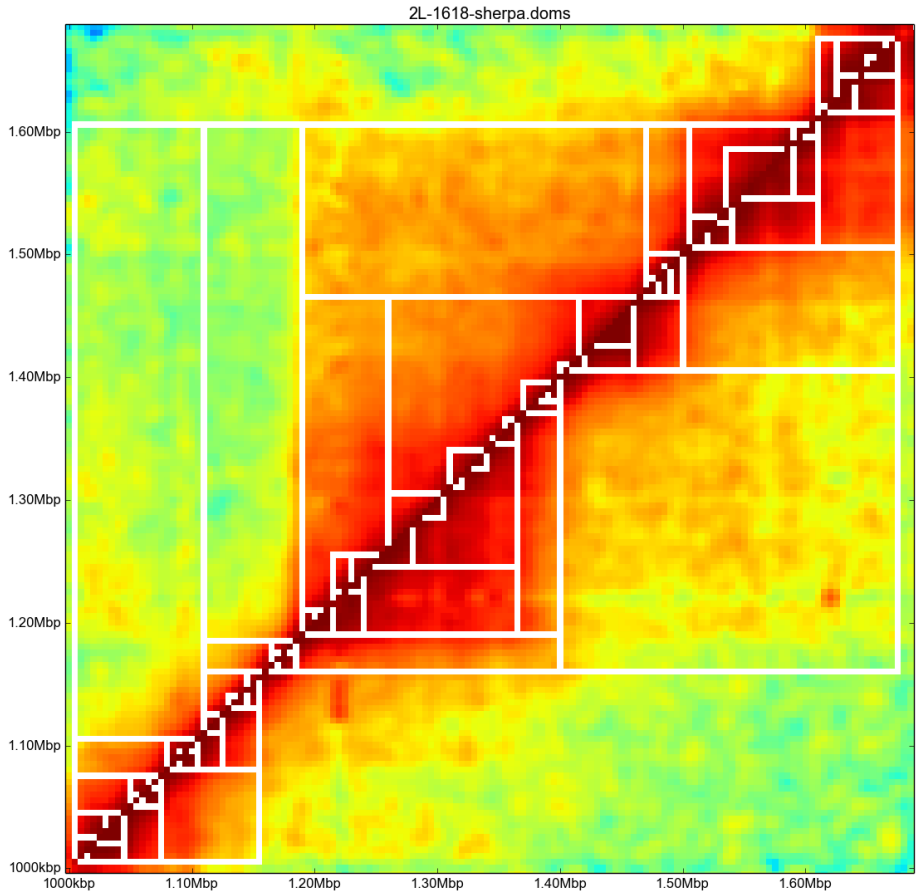
- Having a hierarchy of domains greatly helps in identification of enriched domain-domain contacts
- Using the domain structure the problem of long range contact identification is reduced to statistical hypothesis testing



Cell-type specific contacts on human chromosome 17, Niskanen et al, unpublished

Identification of domain hierarchy rearrangements between cell types

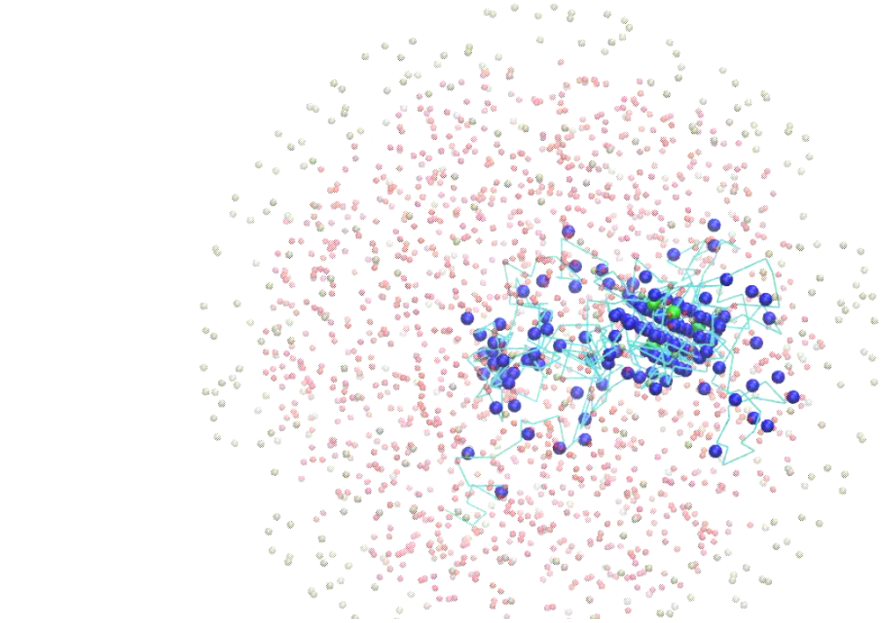
- Identification of hierarchy allows for identification of differences in neighboring domain association between cell types
- This is very frequent at the sub-tad level as opposed to the rather infrequent neighboring „super-TAD” rearrangements
- One needs to define a distance metric on the domain hierarchies to be able to measure the amount of change between conditions

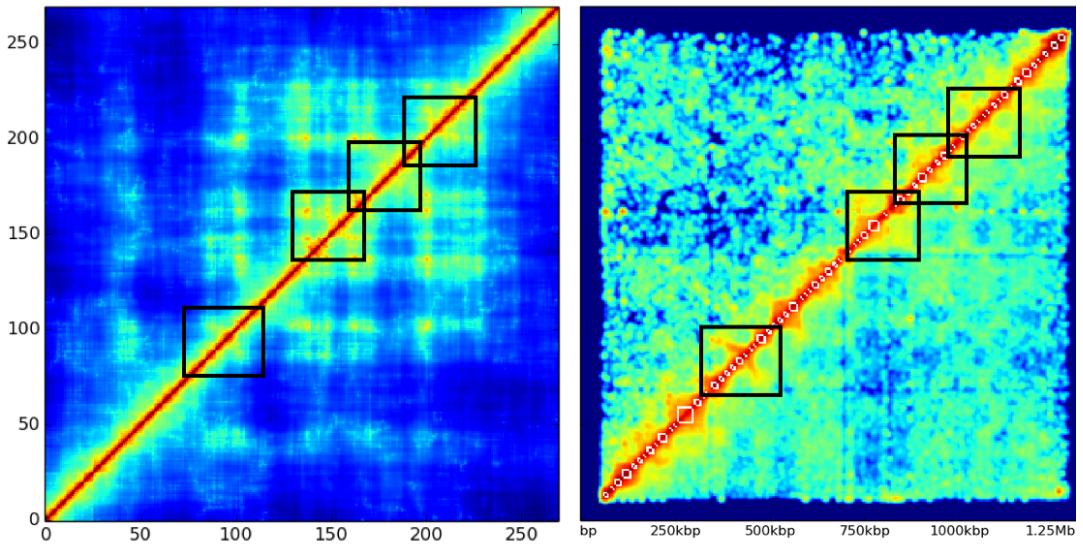


Different emryonic stages in D. melangoaster, unpublished

3d modelling to approach understanding?

- We also develop models and use Monte Carlo simulations to understand the role differen assumptions on the resulting conformation distributions

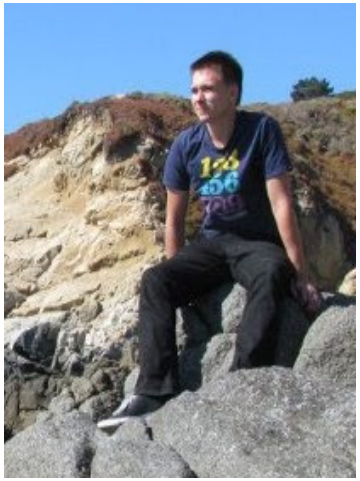




Summary: a XXI century problem?

- Interesting (non-typical) computational problem: large, but finite scale of data, not necessarily a limit problem
- Very quickly developing field on the experimental side (in the last 5 years the matrix size grew a milion times)
- Fundamental (for biology) problems of cell state representation constraints
- Impossible to tackle by imaging alone, need for computational solutions
- Some partial solutions (domain structure, long range contacts, etc) are now available

- Paweł Bednarz
- Julia Herman-Iżycka
- Ilona Grabowicz
- Hania Kranas
- Krzysiek Królak
- Ania Macioszek
- Magda Machnicka
- Karolina Sienkiewicz
- Piotrek Śliwa
- Irina Tuszyńska
- Rafał Zaborowski
- Bartek Zawalski



- Eileen Furlong (EMBL)
- Mara Dierssen (CRG)
- Minna Kaikonnen (UEF)
- Carla Margulies (LMU)
- Norbert Dojer, Ewa Szczurek, Jerzy Tiuryn (MIM UW)

The National Centre for Research and Development

neuron Network of European Funding for Neuroscience Research

EMBO

NARODOWE CENTRUM NAUKI

INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY

FNP Foundation for Polish Science

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND

JEDI